Power Consumption Reduction Effort for the K computer

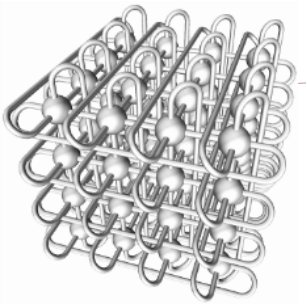
November 16, 2015
RIKEN AICS
Kazuo MINAMI

Copyright 2015 FUJITSU LIMITED

Contents

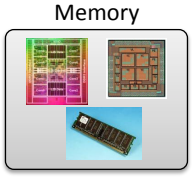
1. To avoid exceeding of power consumption limit
2. Analysis of correlation with power consumption and application profiles
3. A trial of power-saving and operational improvement using automatic collection system of job information

Copyright 2015 RIKEN



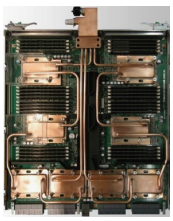
K computer

Node
CPU × 1
ICC × 1
Memory



128 GFLOPS
16 GB

System Board (SB)
(500mm × 500mm)
Node × 4



512 GFLOPS
64 GB

Compute Rack
(800mm × 800mm)
SB × 24
IOSB × 6

12.3 TFLOPS
1.5 TB

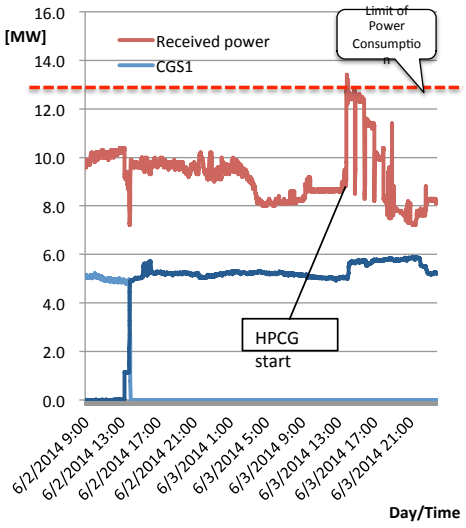
Racks
(4000mm × 2400mm)
Compute Rack × 8
Disk Rack × 2

98.4 TFLOPS
12 TB

Full System
(40m × 40m)
Compute Rack × 864
Disk Rack × 216

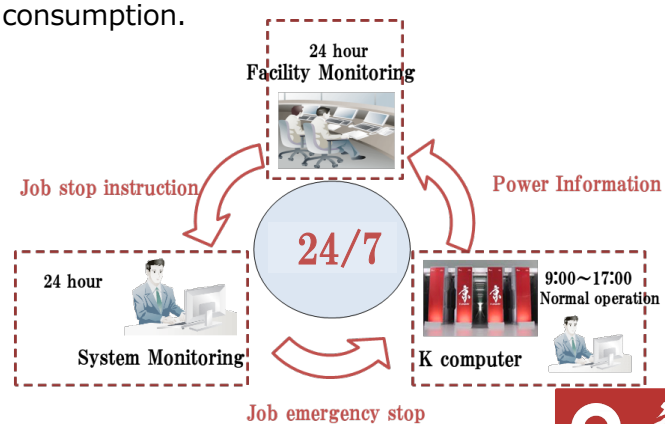
10.6 PFLOPS
1.27 PB

To avoid exceeding of power consumption limit



- Recently jobs with large power consumption have increased and total power consumption exceeded the power consumption limit regulated by contract with provider.
- If the total power consumption exceed the power limit, we have to pay additional charge as a penalty.
- When the excess of power limit occurs, we have to stop some jobs quickly to reduce the power consumption.

Emergency job stop procedure



To avoid exceeding of power consumption limit

- How do I choose the job to stop ?
  - It was decided to select a job to be stopped by estimating the power consumption of each job.
  - This method is a little difficult, but we can stop the proper job.

Estimation Method for Power Consumption by Job.

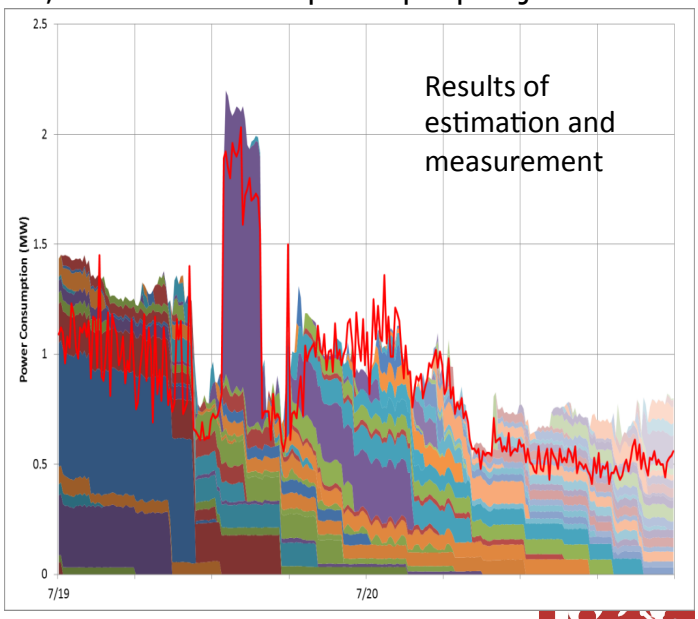
$$P = a \cdot Tcpu + b \cdot Tair + c$$

- CPU water-cooled, Memory air-cooled

P : Power consumption (MW)

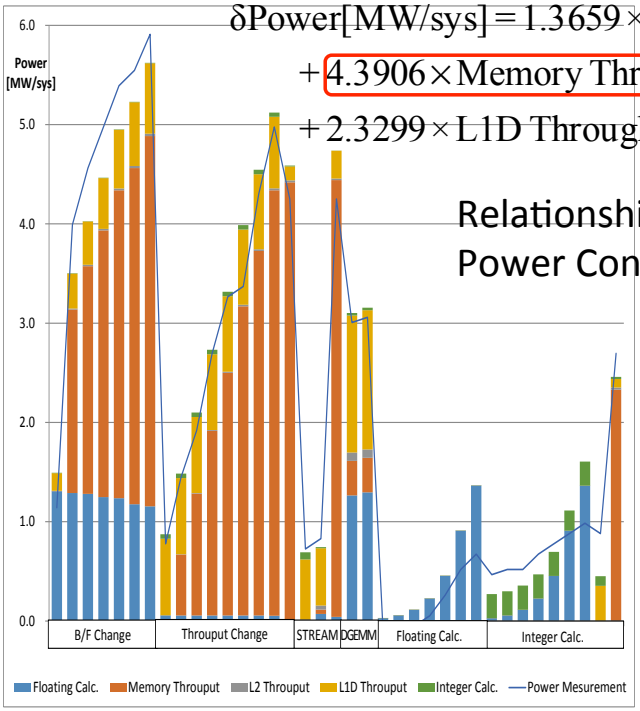
Tcpu : CPU temp. (℃)

Tair : SB Exhaust gas temp. increments (℃)



Analysis of Correlation with Power Consumption and Application Performance

$$\delta Power[MW/sys] = 1.3659 \times Floating\ Calc.[\%] + 4.3906 \times Memory\ Throughput[\%] + 0.0857 \times L2\ Throughput[\%] + 2.3299 \times L1D\ Throughput[\%] + 0.2429 \times Integer\ Calc.[\%]$$

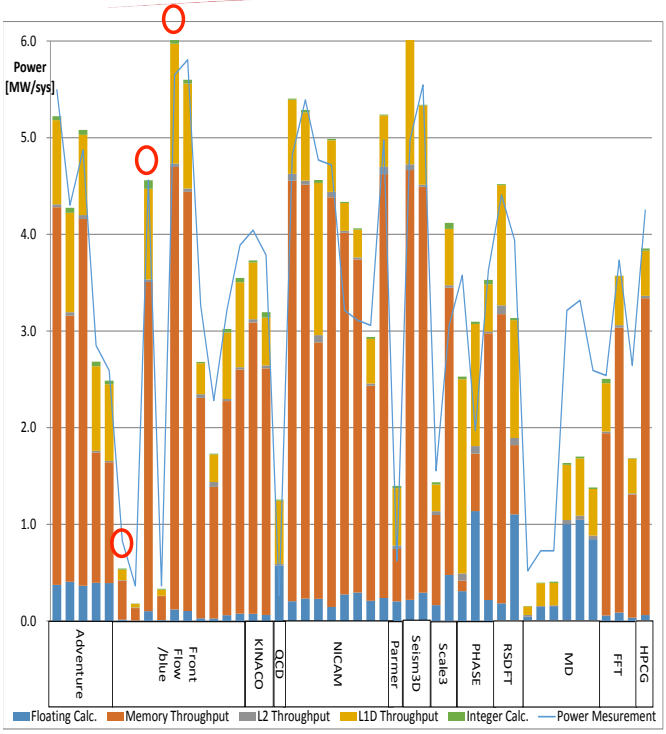


Relationship Each Performance Items and Power Consumption increments

- Memory Throughput has a Greater Impact on the Power Consumption of the K computer.

Simple kernel Loops

Analysis of Correlation with Power Consumption and Application Performance



Real Application Kernel Loops

- We evaluated the power consumption increments of the real application kernel loops using the basic loop formula.
- The estimated power consumption can well reproduce the measured one
- The three red circles correspond to tuned results of the a application
- →Power consumption increments depends on performance tuning status drastically.

A trial of improving of power consumption and system operation

- A extract result of the number of jobs and the node times of production jobs.
  - Nodes > 4rack(384node).
  - Elapse times > 2h.

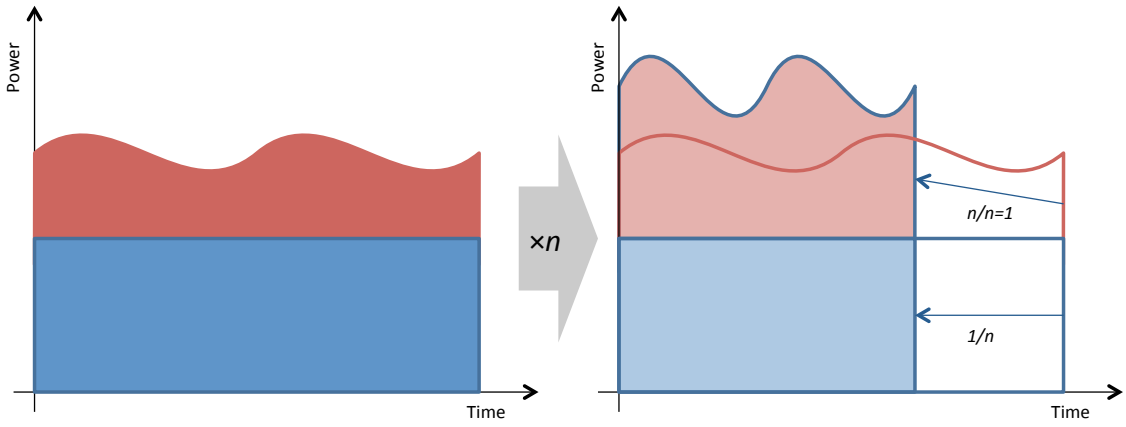
Job Type	# job	# node×time
All Jobs	245,610	260,195,662
Product Job	18,837	179,384,138
Ratio[%]	7.67%	68.94%

- Only the floating point calculation and the memory performance can collect automaticaly using current system.
- A new evaluation formula considering only two performance items is adopted.

$$\text{Power[MW/sys]} = 8.8128 + 1.7895 \times \text{Floating Calc.[\%]} + 4.9921 \times \text{Memory Throughput[\%]}$$

A trial of improving of power consumption and system operation

- A Evaluation Model of Power Consumption of Each Performance tuned Applications.



A trial of improving of power consumption and system operation

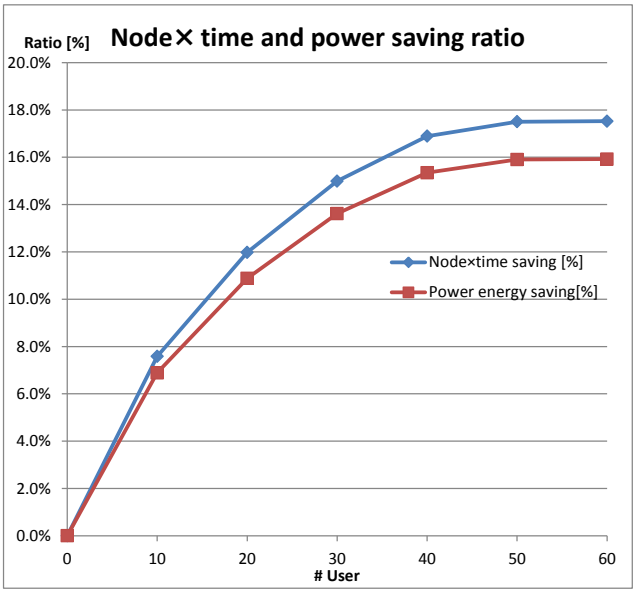
The result of each performance items of each application

					Power energy saving [MWh]		
Field	# job	# node×time	Flops [%]	Memory [%]	+10%	+20%	+50%
Material	37,080	57,923,014	9.64%	13.09%	559	1,026	2,051
Environment	80,757	53,023,661	5.05%	24.64%	512	939	1,878
Engineering	38,187	50,764,035	2.97%	13.02%	490	899	1,798
Physics	23,910	47,021,305	12.98%	18.87%	454	833	1,665
Life	32,150	37,215,166	3.16%	4.79%	359	659	1,318
System	32,097	9,608,454	34.91%	6.18%	93	170	340
Mathematics	1,429	4,640,027	4.26%	8.04%	45	82	164
Sum./Ave.	245,610	260,195,662	7.92%	14.94%	2,513	4,608	9,215

- Performance characteristics was observed in each field.
  - The stencil calculation used a lot of memory throughput.
  - MD and DGEMM calculation used a lot of CPU performance.
- These field average performance is not sufficiently enough high. Simply,
  - With a 10% performance improvement, We can use more 23M node x hours.
  - With a 50% performance improvement, We can use more 87M node x hours.
- If these estimations convert to energy consumption,
  - With a 10% performance improvement, we can achieve energy savings of 2.5GWh.
  - With a 50% performance improvement, we can achieve energy savings of 9.2GWh.
- As the performance evaluation axis,
- We uses a weighted average node time product for all jobs.

A trial of improving of power consumption and system operation

- The performance of each big user's applications has estimated how much differs compared to the field average.
- If both the CPU performance and the memory performance improved to field average, NODE×TIME can be reduced as below.
  - When we improve performance about the 10(50) major user application, we can use more 20M(46M) NODExTIME.
- The power consumption is
  - When we improve performance about 10(50) major user application, we can achieve energy savings of 2.1GWh(4.8GWh).



We are trying to improve power consumption and system operation by using this evaluation.

conclusion

1. To avoid exceeding of power consumption limit
   - ➢ We build the emergency job stopping system that predict power consumption by 30 minutes average.
2. Analysis of correlation with power consumption and application profiles
   - ➢ We formulated the relationship between power consumption increments and each application characteristics.
3. A trial of power-saving and operational improvement using automatic collection system of job information
   - ➢ When we improve performance about 10(50) major user application, we can achieve energy savings of 7%(16%).
   - ➢ We are trying to improve power consumption and system operation by using this evaluation.